REMARKS

Claim 1 has been amended to clarify the subject matter regarded as the invention. Claim 4 has been canceled as duplicative of claim 3. Claims 6-12 were allowed in the parent case and have been canceled from the present application. Claims 13-21 were subject to a restriction requirement in the parent case and are canceled from the present application without prejudice. New claim 22 is added. As a result, claims 1, 2, 3, 5, and 22 are pending.

THE REJECTIONS

In an Office Action mailed March 12, 2001, in the parent application to the present application, the Examiner rejected claims 1-5 under 35 U.S.C. § 103(a) as being unpatentable over Brendel (U.S. Patent No. 5,774,660) in view of Caccavale (U.S. Patent No. 5,459,837).

The rejection is respectfully traversed. With respect to claim 1, Brendel teaches distributing website content among different servers and assigning connections based on which server(s) has/have the requested content. Brendel col. 5, line 47 – col. 6, line 50. Brendel further mentions load balancing based on the number of requests being handled by each server. Brendel col. 9, lines 30-32; col. 12, lines 33-37. Caccavale teaches sending probes to remote clients, the probes being operable to gather response time data at the remote client by sending test requests from the remote client to one or more servers at various times and then sending the response time data back to a central location for analysis. Caccavale col. 1, line 60 – col. 3, line 31.

Claim 1 recites that the system is "configured to monitor connections established between the plurality of servers and clients on the external network" and that the "predictive responsiveness indicators" are "operative to predict the response time of each of the plurality of servers based at least in part on response time data gathered at the system in the course of

monitoring connections established between the plurality of servers and clients on the external network." Neither Brendel nor Caccavale teaches predicting the response time of servers "based at least in part on response time data gathered at the system in the course of monitoring connections established between the plurality of servers and clients on the external network," as recited in claim 1. As such, claim 1 is believed to be allowable.

Claims 2, 3, and 5 depend from claim 1 and are believed to be allowable for the same reasons described above.

THE NEW CLAIM

Similar to claim 1, claim 22 recites "means for gathering response time data at the system in the course of monitoring connections between the plurality of servers and clients on the external network" and "means for predicting the response time of each of the plurality of servers based at least in part on response time data gathered at the system in the course of monitoring connections established between the plurality of servers and clients on the external network." As a result, claim 22 is believed to be allowable for the same reasons described above.

AMENDMENTS TO THE SPECIFICATION

On page 1, beginning at line 1:

[SYSTEM AND METHOD FOR DISTRIBUTING CONNECTIONS TO A GROUP OF MACHINES] SYSTEM FOR DISTRIBUTING LOAD OVER MULTIPLE SERVERS AT AN INTERNET SITE.

On page 1, beginning at line 5:

This application is a continuation in part of co-pending application Serial No. 08/552,807 [(Attorney Docket No. CISCP002)], which is incorporated herein by reference for all purposes.

On page 1, beginning at line 8:

This application is related to co-pending application Serial Nos. 08/850,730 and 08/850,836, [_____, (Attorney Docket Nos. CISCP007 and CISCP008, respectively)] filed concurrently herewith and subsequently issued as U.S. Pat. No. 6,061,349 issued May 9, 2000 and No. 6,104,717 issued August 15, 2000, respectively, which are incorporated herein by reference for all purposes.

On page 12, beginning on line 19:

Requests to internet site 100 from external sites on Internet 102 are routed through Local Director 110. Local Director 110 determines which server of group of TCP based servers [4] 112 should receive the request. Briefly, it does this as follows. A virtual IP address is defined for internet site 100. This virtual IP address is the IP address which the outside world, including the rest of the Internet 102, uses to access internet site 100. To an outside client, it appears that a

single virtual machine hav[redacted] a single virtual IP address services int[redacted]et site 100. The individual identities and IP addresses of the individual servers within the group of TCP based servers 112 are not evident to the user. In certain embodiments, group of TCP based servers [4] 112 may implement more than one virtual server as described in U.S. Patent No. 6,061,349 [application Attorney Docket No. CISCP007 filed concurrently herewith], which is incorporated herein by reference. In such embodiments, a plurality of virtual machines are implemented on different port numbers of a set of real or physical machines. In accordance with the present invention, each virtual machine may allocate connections to a plurality of physical machines.

On page 14, beginning at line 8:

Local Director 110 thus operates to distribute packets among group of TCP based servers 112 by intercepting each packet sent to a virtual machine at internet site 100 and changing the destination IP address in the packet from a virtual IP address which corresponds to all of internet site 100 to a real IP address which corresponds to a single physical machine located at internet site 100. In certain embodiments, Local Director 110 includes more than one virtual machine IP address and therefore routes connections for more than one virtual machine to a set of physical machines through the physical machine's ports. Additionally, in such embodiments, it is also possible that each physical machine is mapped to more than one virtual machine. Such a system is described in detail in co-pending application Serial No. 08/850,730, now U.S. Patent No. 6,061,349, [_____, (Attorney Docket No. CISCP007) filed concurrently herewith, and] previously incorporated by reference.

On page 15, beginning at line 1:

By adopting a single virtual IP address for the entire server group of TCP based servers 112, the problems of round robin DNS and DNS caching are avoided. Specifically, any

connection made to the virtual IP address of a virtual machine is perceived by the connecting entity as a connection to the virtual machine and not as a connection to the physical machine to which the connection is physically made. The connecting entity never discovers the real IP address of the real machine handling the connection since, for outgoing packets, the real machine source IP address is replaced with the virtual machine IP address by the Local Director. Therefore a connecting entity which caches IP addresses using DNS caching caches the virtual machine IP address and will not address connection requests exclusively to any one server from group of TCP based servers [4] 112 to the exclusion of the other servers in group of TCP based servers [4] 112.

On page 17, beginning at line 15:

In certain embodiments, servers are failed when they fail to make a certain number of consecutive connections corresponding to a failure threshold. In some embodiments, each failed connection itself fails only after repeated attempts to make the connection are unsuccessful. The predicted response time for such a machine would still match the aged predicted response time from its last successful response. That predicted response time would be unduly optimistic since if the server has failed, then, in fact, the actual response time is going to be at least as long as it takes to fix the machine and bring it back up on line. The selection of the server based on its unrealistically good response time is therefore overridden by a failure flag. A further description of a system in which failures of individual physical machines are determined and failed machines are tested to determine if they can be placed on line again is described in detail in co-pending application Ser. No. 08/850,836, now U.S. Pat. No. 6,104,717, [Attorney Docket No. XXX1-P0XX] filed concurrently herewith, which is incorporated herein by reference for all purposes.

1. (Amended) A system for distributing connections from clients on an external network to a plurality of servers on an internal network, the system comprising:

a client interface to the external network the client interface being operative to receive and send packets to and from a remote client;

a server interface to the internal network, the server interface being operative to receive and send packets to and from a plurality of servers, the plurality of servers being operative to establish a connection with the remote client and the system being configured to monitor connections established between the plurality of servers and clients on the external network;

a plurality of predicted responsiveness indicators, each of the plurality of predicted response indicators being associated with [each] at least one of the plurality of servers, the predicted responsiveness indicators being operative to predict the response time of each of the plurality of servers based at least in part on response time data gathered at the system in the course of monitoring connections established between the plurality of servers and clients on the external network, the predicted responsiveness indicators also being stored within the system in a manner that the predicted responsiveness indicators may be accessed; and

a predicted responsiveness comparator which is operative to access and compare the predicted responsiveness indicators and to determine which servers from among the plurality of servers is associated with a predicted responsiveness indicator which measures a best response time, the predicted responsiveness comparator being further operative to select a pointer to a server which has a predicted responsiveness that is the best predicted responsiveness among the predicted responsiveness of the plurality of servers.

whereby the server which has a predicted responsiveness which is the best predicted responsiveness is selected to handle the next connection from a client.

Reconsideration of the application and allowance of all claims are respectfully requested based on the preceding remarks. If at any time the Examiner believes that an interview would be helpful, please contact the undersigned.

Respectfully submitted,

William J. James
Registration No. 40,661
V 650 903 3502
F 650 903 3501

VAN PELT AND YI, LLP
4906 El Camino Real, Suite 205
Los Altos, CA 94022